PROYECTO VERIFACTURA

PROCESAMIENTO DE FACTURAS VEHICULARES ECUATORIANAS

Grupo 1

Andrea Fernanda Morán Vargas

Pedro José Vidal Orús

Contexto

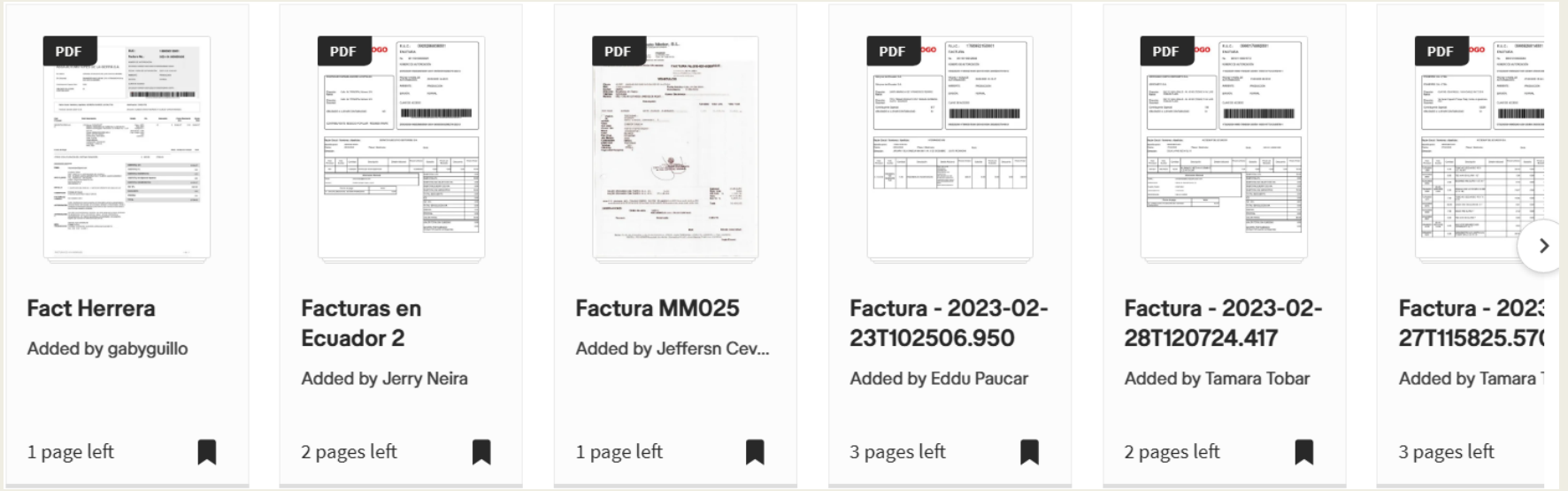
- Las instituciones financieras, concesionarias y entidades tributarias manejan miles de facturas vehiculares con estructuras heterogéneas.
- La digitalización ha generado documentos en PDF escaneados y nativos, pero la información crítica sigue dispersa.
- En la práctica se presentan con formatos muy variados: algunas provienen de sistemas nativos con texto embebido, mientras que otras son simples escaneos sin estructura digital.
- Esto genera múltiples problemas para instituciones financieras y organismos de control, que necesitan procesar la información de forma ágil y confiable.

Proyecto

- El proyecto tuvo como objetivo principal la construcción de un pipeline automatizado para la extracción de información de facturas vehiculares.
- En su dimensión académica, debía cumplir con los componentes clásicos de un curso de machine learning: exploración de datos, limpieza, ingeniería de características, balanceo, aumento de datos, particionado y construcción de un pipeline de preprocesamiento.
- Pero además, se buscaba responder a una necesidad práctica, orientando el resultado hacia un sistema documental capaz de procesar documentos reales y entregar salidas en JSON que cumplieran con estándares de calidad, auditabilidad y trazabilidad.

Dataset

- El punto de partida fue un conjunto de facturas obtenidas desde Scribd y otras fuentes abiertas. Estas facturas fueron transformadas en un archivo CSV que consolidó alrededor de treinta variables, entre las que destacaban campos como CLASE, MARCA, MODELO, MOTOR, SUBTOTAL, IVA y TOTAL.



Exploración Inicial

- El análisis exploratorio inicial reveló problemas importantes: una gran proporción de valores faltantes en variables técnicas como MODELO HOMOLOGADO ANT, formatos inconsistentes en fechas y montos, y un fuerte desbalance de clases en la variable objetivo. Estos hallazgos justificaron la necesidad de implementar rutinas de limpieza antes de cualquier intento de modelado.

FECHA_DO	DIRECCIOI	MODELO_	SUBSIDIO	AÃ'O	SUBTOTAL	CLASE	TOTAL	CILINDRAJ	MODELO	MODELO_	RAMV_CP	RUEDAS	DESCUENT	NUMERO_	COLOR	MOTOR	NOMBRE_	CAPACIDA	MARCA	RUC	COMBUST	EJES	TIPO	IVA	CONCESIC	TONELAJE	VIN_CHAS	PAIS_ORIG
########	Dir. Matriz: AVENIDA ATAHUALPA SN y RIO GUAYLLABAMBA																											
########	Av. de los Granados E11-67 y La			2009	25810.96	CAMION	30400	3900 C.C.	HD72 CHASIS CABIN/		T00715845		0	010-001-0	BLANCO	D4DB8357	UNIDAD D	5 PASAJER	HYUNDAI	1,79E+12	DIESEL		CAMION	1331.9	HYUNDAI	2.8	KMFGA17I	ECUADOR
########	Av. Indoamerica Km 3 Â½			2022	20535.71	CAMION	23000	2771	HOWO ZZ1047D3414		T02691620		437.5	001-801-0	BLANCO	35121437	AGROCELHONE DEL I		SINOTRUK	1,89E+12	DIESEL		CAMION	2464.29	VEHICENT	3.6	LZZ5BADC	CHINA
########	AV.32 N.O. Solar 37 - MZ 1, Av.			2016	35086.84	CAMIONE	39999		AMAROK BI-TDI 2H	BA34 AC 2.0 CD 4X4 TM			0	001-101-0	PLOMO	CNE08628	TRADING BROKERAG		VOLKSWA	9,93E+11	DIESEL		DOBLE CA	4912.16	INDUWAGEN S.A.		WV1ZZZ2H	ZGA02743
########	Av. 10 de ¿D-MAX CRDI 3.0 CD 4			2014	28003.57	CAMIONE	31364	2999	D-MAX CR	D-MAX CRDI 3.0 CD 4X2 TM DIE			0	002-100-0	PLATEADO		RUIZ CON	5	CHEVROLE	1,79E+12	DIESEL		DOBLE CA	3360.43	AUTOMOT	1.25		
########	Av. 10 de Agosto N45-266 y Av.			2016	15982.14	AUTOMOV	17900	1398	SAIL AC 1.4 4P 4X2 TI		B7740044	4	0	002-100-0	PLOMO	LCU16033	QUIZHPI JI	5	CHEVROLE	1,79E+12	GASOLINA	2	SEDAN	1917.86	AUTOMOT	0.37	8LAUY527	ECUADOR
########	Matriz: Av LOGAN EXPRESSION			2018	15816.47	AUTOMOV	17714.45	1598	LOGAN EX	LOGAN EX	T0215890	4	0	034-010-0	PLOMO	K7MA812I	TERRANO\	5	RENAULT	1,79E+12	GASOLINA	2	SEDAN	1897.98	AUTOMOT	1.54	9FB4SREB	COLOMBIA
########	Av. Mariana de JesÃºs Oe3-283			1998	14098.21	AUTOMOV	15790	1498	CAVALIER STD 2.2 2P		B7740113833		1964.29	004-101-0	TINTO	24576436	DIAZ BATI	5	CHEVROLE	1,79E+12	GASOLINA		PASAJERO	1691.79	AUTOLANI	1.325	3G1JX144	MEXICO
########	Av. 10 de ¿D-MAX CRDI HI RIDE			2022	27677.68	CAMIONE	30999	2499	D-MAX CR	D-MAX CR	B7740142	4	0	001-100-0	VINO	4JK1 WS3	BALCAZAR	5	CHEVROLE	1,79E+12	DIESEL	2	DOBLE CA	3321.32	AUTOMOT	1.25	8LBETF3P4	ECUADOR
########	AV. RIO COCA E8-73	0		2025	399.13		459		TRAILBLAZER HIGH COUNTRY AC 2.8 5P 4X4 TA DIESE					016-108-0	NEGRO		IMPORTADORA VIDRIALUM S.A			1,79E+12	DIESEL			59.87	ECUA-AUTO S.A. ECA		9BG156PK0	SC422166
########	PASAJE A 8 y CALLE 6	0			26		26		HILUX			4		001-901-000000636			GUARNIZO AGUILAR	TOYOTA		1,71E+12		2	PICKUP	0	TALLERES DIESEL Y GASOLINA			
2/7/2023	VÃ-a a Daule Km 7.5	0			514.02		575.7							008-013-000003384			ACCEQUIP DEL ECUADOR S.A.			9,91E+11				61.68	FEHIERRO CIA. LTDA.			
########	KM 11.5 VIA A DAULE	0			73.39		82.2							005-011-000019772			ACCEQUIP DEL ECUADOR			9,90E+11				8.81	GERONIMO ONETO GERONETO S.A.			
2/1/2024	AV JUAN TANCA MAR	0			4.46		5							271-021-000034178			AGROCELHONE DEL ECUADOR (			9,91E+11				0.54	SERVIENTREGA ECUADOR S.A.			
########	Calle AV. PRINCIPAL S	0			10		10							001-100-000000029			SERVICIO EJECUTIVO SERTURIS			9,20E+11				0	RODRIGUEZ MIRABA ANDRES LEOPOLDO			
########	ELOY ALFARO N28-16	2.67			36.65		42.15							098-008-001040088			INMOCISNE CIA LTDA			1,79E+12	EXTRA			5.5	AUTOMOVIL CLUB DEL ECUADOR			ANETA
########	Barrio Puerto Nuevo	0			66.96		75							001-100-000000019			SERVICIO EJECUTIVO SERTURIS			9,17E+11				8.04	VALDEZ LEON OSWALDO JOSE			
########	ELOY ALFARO 218 Y E	0			58.04		65							072-002-000004554			ANGEL VINICIO HEREDIA MAYO			1,79E+12				6.96	AUTOMOVIL CLUB DEL ECUADOR			ANETA
########	Catalina Aldas y Port	0			0.88		0.99							001-003-015728988			AGROCELHONE DEL ECUADOR (			1,79E+12				0.11	DELIVERY HERO DH E-COMMERCE			ECUAD

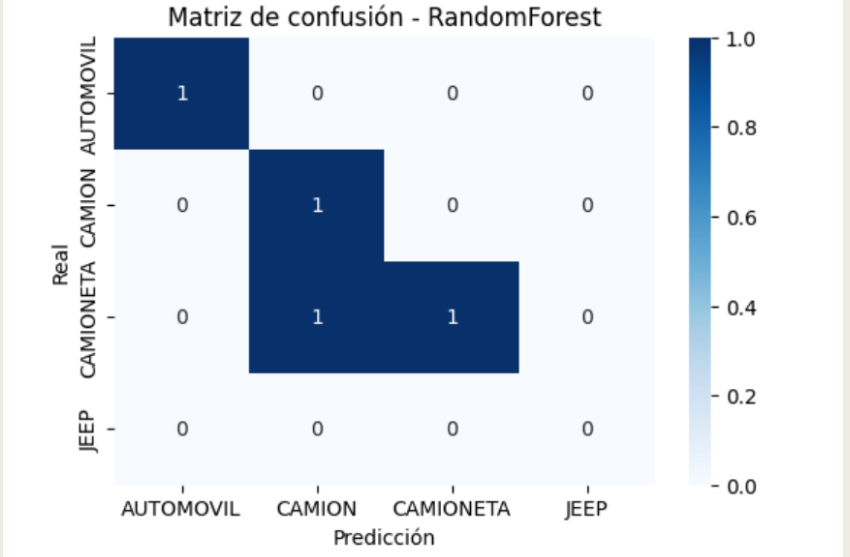
Modelado Tabular y Limitaciones

- Con el dataset preparado, se entrenó un modelo RandomForest para predecir la clase de vehículo. Los resultados iniciales mostraron un accuracy de 0.75 y un F1-macro cercano a 0.78.
- La confusión recurrente entre CAMIONETA y CAMIÓN evidenció que la información tabular era insuficiente para capturar diferencias clave que, en los documentos originales, estaban ligadas al layout y a descripciones contextuales.

Accuracy: 0.75
F1 macro: 0.7777777777777777

Reporte:

	precision	recall	f1-score	support
AUTOMOVIL	1.00	1.00	1.00	1
CAMION	0.50	1.00	0.67	1
CAMIONETA	1.00	0.50	0.67	2
accuracy			0.75	4
macro avg	0.83	0.83	0.78	4
weighted avg	0.88	0.75	0.75	4



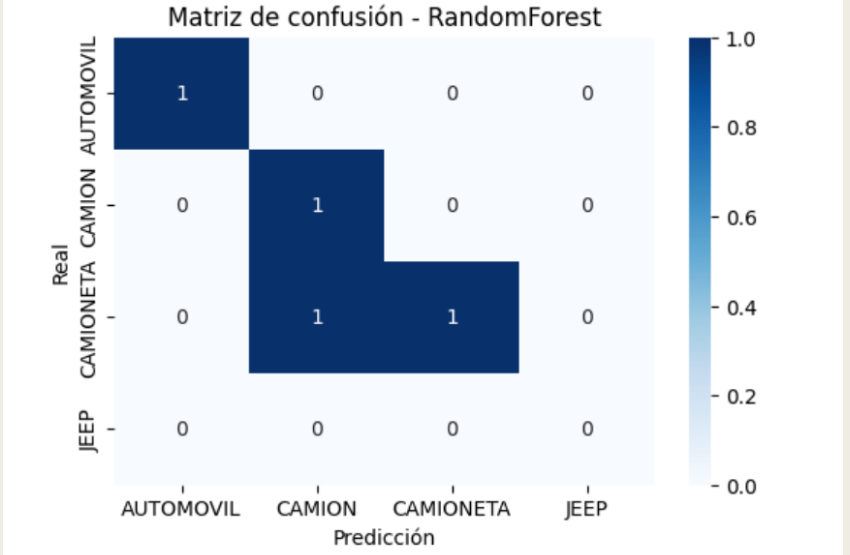
Punto de inflexión

- Etamaño reducido del dataset y el desbalance de clases imposibilitaban una partición estratificada robusta.
- Estos hallazgos marcaron un punto de inflexión, al dejar en claro que el enfoque tabular no resolvía adecuadamente el problema.

Accuracy: 0.75
F1 macro: 0.7777777777777777

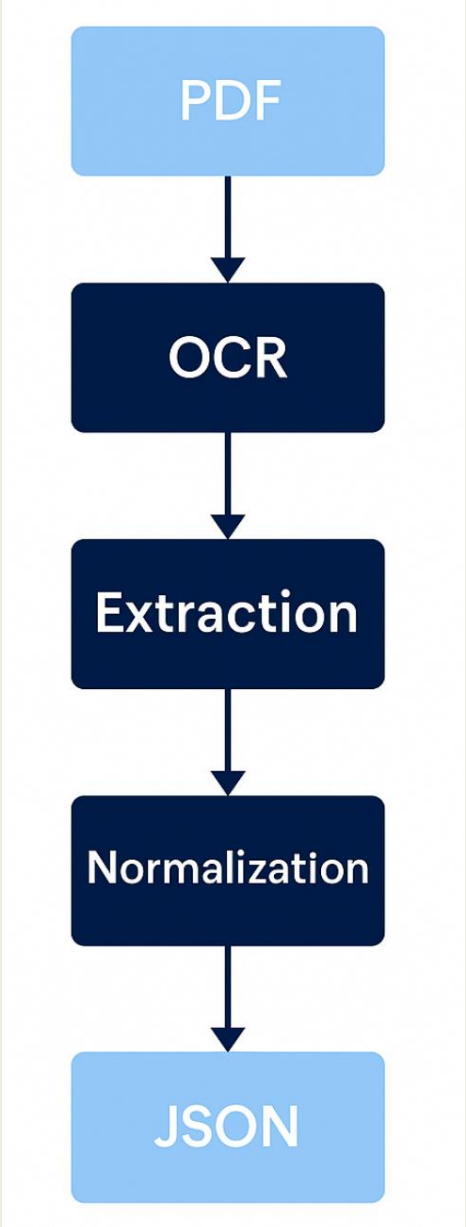
Reporte:

	precision	recall	f1-score	support
AUTOMOVIL	1.00	1.00	1.00	1
CAMION	0.50	1.00	0.67	1
CAMIONETA	1.00	0.50	0.67	2
accuracy			0.75	4
macro avg	0.83	0.83	0.78	4
weighted avg	0.88	0.75	0.75	4



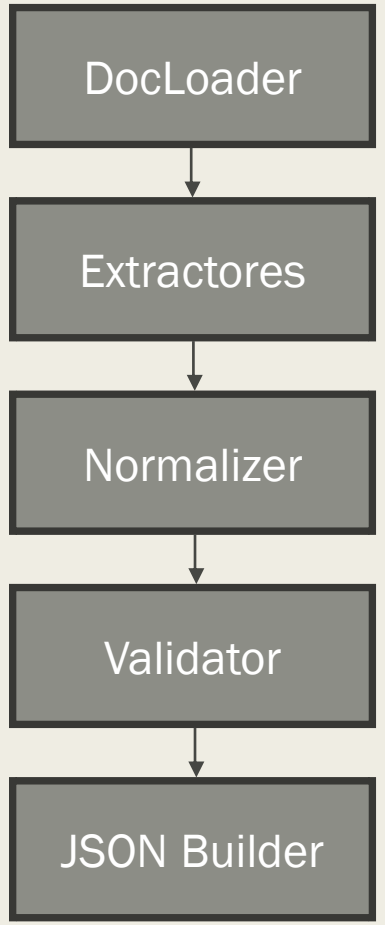
Cambio de Paradigma

- Ante las limitaciones del enfoque tabular, el proyecto evolucionó hacia un paradigma documental.
- Esto implicó reconocer que una factura no es una fila en un CSV, sino un documento estructurado visualmente, en el que cabeceras, tablas y paneles de totales cumplen funciones semánticas específicas.
- Por ello, se propuso un pipeline basado en la lectura de PDFs, integración de OCR para documentos escaneados, extractores guiados por reglas o plantillas, normalización de formatos y validación de consistencias.



Arquitectura del Pipeline Documental

- La arquitectura del pipeline documental se concibió como un paquete Python modular denominado verifactura_pipeline.
- Este incluye distintos componentes que interactúan de manera secuencial.
  - **El DocLoader** abre los documentos y extrae el texto.
  - **Los extractores** genéricos o específicos localizan los campos de interés.
  - **El Normalizer** homogeneiza formatos.
  - **El Validator** aplica reglas de negocio para verificar consistencias, como la relación entre SUBTOTAL, IVA y TOTAL.
  - **El JSONBuilder** genera la salida estandarizada.



Roadmap y Futuro

- El pipeline desarrollado constituye apenas el primer paso de un proyecto con gran potencial de evolución.
- Se contempla integrar un OCR robusto como Tesseract o Azure OCR, capaz de manejar documentos escaneados con baja calidad.
- También se planea ampliar la validación, incluyendo el checksum de VIN y la verificación de RUC con módulo-11.
- Se construirá un dataset con mayor número de facturas y realizar un dataset JSON(L) de resultados de extracción, lo que abrirá la puerta a realizar un finetune sobre modelos LLM.

Conclusiones

- El proyecto *Verifactura* permitió aprender de manera práctica que no todos los problemas con datos pueden resolverse mediante modelos tabulares.
- El enfoque inicial fue útil para cumplir con los objetivos académicos, pero resultó insuficiente para capturar la naturaleza documental del problema.
- La transición hacia un pipeline documental fue el paso clave que alineó la solución técnica con la realidad de las facturas vehiculares.
- El resultado es una arquitectura modular, extensible y auditable, capaz de extraer campos críticos y generar salidas confiables.
- En corto plazo, el proyecto podrá escalar en volumen, integrar OCR y plantillas adicionales, y apoyarse en modelos de IA sofisticados.